Niveau(x) d'enseignement: 4 Numéro EDUQ : 8670169
Centre d'intérêt: 550

Bertrand, Richard G., 1949-.

Components of variance approach for the observation data of the IEA Classroom Environment Project / Richard Bertrand. -- Sainte-Foy : INRS-Education, 1985.
16 f. ; 28 cm.

"Communication présentée à la rencontre annuelle de l'American Educational Research Association, Chicago, 1985". Code : C-140.
Bibliogr.: f. 15-16.

Vedette(s)-secondaire(s): INRS-Education

Descripteur(s): Pratique pédagogique * Observation de la classe * Relations maître-élèves * Secondaire (Ordre d'enseignement)

L'auteur présente dans cette communication les résultats d'une étude de fiabilité qui portait sur les pratiques pédagogiques telles qu'observées en classe. Il s'agit d'une partie d'un projet beaucoup plus vaste dont l'objectif est d'identifier les corrélations existant entre les pratiques pédagogiques et les résultats scolaires. L'observation a porté sur trente classes de mathématiques de septième et huitième année. L'enseignement fut observé à dix occasions, par dix observateurr, à l'aide d'une instrument appelé "Five Minute Interaction" qui permet de mesurer les interactions entre les professeurs et les élèves. Les résultats de l'étude sont assez décevants: beaucoup de variables ne se sont pas avérées fiables.

Type de document: r

P69/85

EDUQ 8670169

COMPONENTS OF VARIANCE APPROACH FOR THE OBSERVATION DATA OF THE IEA CLASSROOM ENVIRONMENT PROJECT

RICHARD BERTRAND

C-140

Communication présentée à la rencontre annuelle de l'American Educational Research Association, Chicago, 1985.

INRS-Éducation
2383 chemin Ste-Foy
Ste-Foy
G1V 1T1

INTRODUCTION

DESIGN

GENERALIZABILITY COEFFICIENT

DATA COLLECTION AND INSTRUMENT

STUDY RESULTS

CONCLUSION

INTRODUCTION

In the next minutes I will talk about the results of a reliability study which had to do with observational data on teaching practices. We conducted this study with grade 7 and 8 mathematics classes, that is about thirteen to fourteen years old students. After the introduction, I will talk about the design of the study, the generalizability coefficient used, the data collection, the instrument and the study results. I will end with a brief conclusion.

This study is part of the IEA Classroom environment research project. Its objective is to identify correlations between teaching practices on the one hand, and educational achievement and pupil attitudes on the other.

This study was conducted because we think it is at least as important to evaluate reliability for the teaching practices as it is for educational achievement. In fact, unreliable measurements cannot be expected to relate to any other measurements. So we will use only reliable teaching practices in order to relate them to educational achievement and pupil attitudes.

Before proceeding with the reliability study, I would like to say a brief word concerning the way we are going to relate the reliable teaching practices to educational achievement.

We will use first ordinary least squares multiple regression to relate each teaching practice to educational achievement that is residual achievement using pre-test and number of hours of algebra as predictors. Second, because we know we have too few data and problems of outliers, we will use resistant line technique of Tukey's type (Velleman and Hoaglin [1981]).

We will then compare the least square fit to the resistant fit and check for curvilinearity and heteroscedasticity.

Now back to the reliability study.

We chose generalizability theory to evaluate the reliability of observational data on teaching practices. This allowed us to measure the importance of different sources of variation in the data: the classes, the observers and the occasions, the last two being error variation.

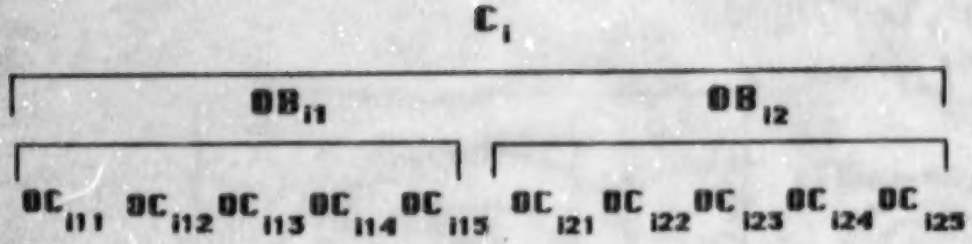
THE DESIGN

This is the observational design we had. Each of the thirty classes was observed by two different observers and each of them observed on five separate occasions. (FIGURE 1)

Visually, this design can be represented by this Venn diagram. (FIGURE 2)

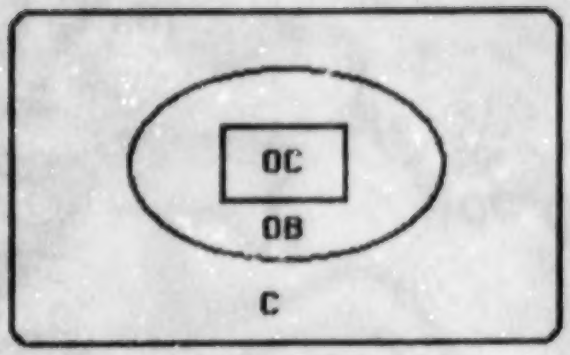
As we can see, the design is entirely nested: the facet occasion (OC) is nested within the facet observer (OB) which in turn is nested within the facet class (C). Frequently, studies of generalizability will include several facets which are crossed, the point of which is to

FIGURE 1



$1 \leq i \leq 30$

FIGURE 2



evaluate interactions such as "class X observer", "class x occasion". We adopted a more simple nested design right from the outset, and this for the following reasons.

First of all, we thought it was not useful to evaluate interactions for this study. In fact, looking at the interaction "class X occasion", given that each class takes place in a different room, it is physically impossible to observe the "same" occasion of several classes. Therefore it was not possible to cross the facet "class" with the facet "occasion" in order to estimate the interaction resulting from "class x occasion". Then looking at the interactions "class x observer" and "occasion x observer", several observers would have to have been in the same class at the same time in order to account for such interactions. To do that would have been both difficult and costly. Now data collected for similar studies indicate that the variation resulting from this type of interaction is minimal when compared to variation due to occasions. Second, considering the problems of sampling errors related to estimates of variance components, we'd like to point out that these problems can become particularly acute when a crossed design with only a few levels per facet is used (Smith 1978).

Finally, in view of this state of affair, time and money lacking, we decided not to use a more complex crossed design.

We further made the assumptions that each facet of this design is considered random. Indeed the thirty classes which were observed are not the only ones which interest us. The same can be said for the observers and the occasions. We wanted to generalize the study results to other similar classes, observers and occasions.

Generalizability coefficient

In order to determine the generalizability coefficient, the contribution of each of the three facets - class, observer, and occasion - to the total variance had to be evaluated. The variance resulting from the differences between classes has been called universe score variance. If the data are to be considered reliable, it is essential that the universe score variance be as large as possible. Error variance, which is what we want to minimize, comes from differences between observers on the one hand, and differences between occasions on the other.

The generalizability coefficient tells us to what extent differences in teaching practices used in each class were reliably observed. The coefficient is represented by this formula:

$$\rho^2_C = \frac{S^2_C}{S^2_C + 1/N_{OB}\, S^2_{OB} + 1/N_{OB}N_{OC} S^2_{OC}}$$

where - the "S" square are estimates of variance components related to classes, observers and occasions;
- N_{OB} is the number of observers per class; and
- N_{OC} is the number of observation occasions per observer, per class.

For each class, we have a fixed number of observers, 2, and a fixed number of occasions, 5. Thus the total number of observation occasions per class is 10.

The variance estimates are obtained in the usual way by means of an analysis of variance with a three-way completely nested design (random model).

DATA COLLECTION AND INSTRUMENT

Thirty math classes of grade 7 and 8 who were being taught basic notions of algebra were observed on ten different occasions. These classes came from thirteen different schools belonging to four school boards of the region of Quebec.

Ten observers were trained in order to master the observation instrument called "Five Minute Interaction" (FMI). The categories of this instrument consist of a series of behaviours linked to teacher - student interaction.

Here is the summary of the codes of this instrument (Table 1). Using the appropriate codes the observer notes, every five seconds, the instructional context, who is speaking, to whom that person is speaking and what is being said. Five of these five-minute observation periods comprise one observation occasion.

STUDY RESULTS

Before presenting the results we must say that we retained three levels of reliability. The first level is the one which corresponds to the limit generally considered acceptable, that is .7. We called it effective reliability.

Table 1 Five-minute interaction: summary of codes

CONTEXT

1- L The teacher is interacting with more than 8 students.
2- S The teacher is interacting with a group of 8 or fewer students.
3- M The teacher is moving around the room while students work independently.
4- T The teacher and the group are in transition.
5- P The teacher is interacting with an individual student outside a group context.
6- N The teacher is not interacting with students.

WHO	TO WHOM
7- T Teacher	14- T Teacher
8- S Student-after first code	15- G Group
9- V Student volunteer	16- S Student
10- N Student non-volunteer	17- I New individual student
11- I Student initiates/callout	
12- C Group callout or no-one	
13- G Group together	

WHAT - ACADEMIC

Instruction	Response
18- Le Lectures/explains	28- Re Response
19- Ex Gives example	29- Rc Recite
20- Sr Structuring	30- St Statement
21- Pb Probe	31- Dk Don't know
22- Aa Alerts as to accountability	
Questions	**Feedback**
23- Rr Requests response	32- Ac Acknowledgement, praise (with P)
24- Hq High-level question	33- Ra Repeats answer
25- Rq Recall or recognition question	34- Rd Redirect
26- Oq Opinion/open-ended question	35- Ga Gives answer
27- Eq Effectiveness question	36- Wr Wrong
	37- Pu Punish

NON-ACADEMIC	OTHER
38- Di Discipline	41- S Silence
39- Pr Procedure	42- Ch Can't hear
40- So Social	

QUALIFIERS	
43- Nv Non-verbal	The code R used in the margin indicates that the preceding sequence is repeated.
44- Em Emphatic	
45- P Positive	
46- N Negative	
47- M Materials	

The second level regroups variables whose coefficient is less than .7 but for which on optimal design has been found.

Now what is an optimal design? When the coefficient does not reach the .7 level based on the design described above, that is 2 observers and 5 occasions, we can calculate a new coefficient, varying the number of observers and occasions, but such that the product of the number of observers and the number of occasions does not exceed ten, the total number of occasions per class in the initial design. For example, we could calculate a new coefficient with $N_{OB} = 3$ and $N_{OC} = 3$, or $N_{OB} = 4$ and $N_{OC} = 2$ or else $N_{OB} = 5$ and $N_{OC} = 2$.

Now if this new coefficient exceeds or equals .7, the corresponding numbers of observers and occasions that were used are chosen for the new optimal design.

This second level refers to what we call potential reliability.

The third level includes variables with a coefficient less than .7 and for with we did not find an optimal design. At this level, we have no reliability.

Here are now the study results.

At the end of data collection, 271 variables (teaching practices) were retained for the reliability study. This means that a generalizability coefficient was calculated for each of these variables. In this way, the reliability of the data for each variable was evaluated.

As can be seen on table 2, of these 271 variables, only 72 were judged reliable at level 1 and 29 at level 2.

Table 2

Level of reliability \ Context	1	2	3	4	5	6	All-context	TOTAL
1. $\bar{\rho}_C^2 \geq .7$	44	0	5	0	11	0	12	72
2. $\bar{\rho}_C^2 < .7$ with optimal design	13	0	5	0	4	0	7	29
3. $\bar{\rho}_C^2 < .7$ without optimal design	75	19	31	13	10	1	21	170
TOTAL	132	19	41	13	25	1	40	271

None of the variables from contexts 2, 4 and 6 reaches a reliability level of 1 or 2. However it must be pointed out that these contexts represent very few variables. As well, 33% of the variables from context 1, 12% from context 3, and 44% from context 5 reach level 1 reliability. In total 27% of the 271 variables are classed in the first level of reliability and 11% in the second. Such low percentages may appear unusual. However they are in line with results from other similar studies (Berliner, 1980 ; Shavelson and Atwood-Russo, 1977 ; Erlich and Shavelson, 1978 ; Erlich and Borich, 1979, Calkins et al, 1977).

Now if we analyse the results by category of teaching activity, we find that the subcategories questions, responses and feedback contain proportionally more reliable variables than the subcategory instruction

or the non academic, uninvolved or other categories. It should be noted that the instruction subcategory, which includes some variables with very high frequencies, has very few variables with level 1 reliability, thus highlighting the considerable instability of the variables in this subcategory. On the other hand, at least half the variables in the questions and responses subcategories attain reliability level 1 or 2.

It is possible to compare these results with those of other studies.

Erlich and Shavelson (1978) reported that nine out of ten variables relating to the teacher's presentation of content were considered to be unreliable.

In addition studies by Erlich and Shavelson (1978), Erlich and Borich (1979) and Calkins et al (1977) reported on a certain number of reliable variables related to questions and feedback. These results resemble our own.

Now for the conclusion.

CONCLUSION

It may seem surprising that so many variables were unreliable. However other studies have obtained similar results. Can these poor results be explained? Are they due to methodological and/or theoretical problems?

There is no shortage of methodological questions. For example, was the observational design suitable? Would it not be advantageous to sample observation occasions in such a way as to observe teaching time in a more representative way? However, with respect to this last question,

this may be easier said than done. Bad weather, illness on the part of observers or teachers are among the factors which can disturb plans for observation sessions.

Can a better combination of the number of observers and the number of occasions be found? Theoretically speaking, this is certainly the case. Indeed what the calculation of the optimal design shows is this: had we used a design with five observers and two occasions each instead of two observers and five occasions each, we could have expected an increase of the order of 40% in the number of variables at level 1 reliability. However is this a realistic design? Will teachers allow five different people to come and observe in their classes (even if the five do not observe at the same time)?

We also feel that much work remains to be done on theoretical issues. Were all the facets likely to influence data reliability used? Quite possibly certain facets with an important role in explaining variations between observations were not accounted for. McGaw et al (1972) and Calkins et al (1977) have already raised this possibility. These authors are particularly critical when it comes to leaving out the "situation" facet in the process of choosing an observational design. This facet refers to aspects such as the period in the school year, the day of the week, the time of the day , the subject being taught, the presentation of new content or the revision of an old one. However including such a facet should not unreasonably complicate the observational design. Otherwise the cost of the study is likely to increase excessively.

Although the planning of such studies raises considerable methodological and theoretical problems, we think they must nevertheless be done. Though such studies are costly, this must be done in order to

evaluate the reliability of teaching practices and as a result increase the likelihood of detecting stable relationships between these practices and student achievement and attitudes.

REFERENCES

BERLINER, D.C. Studying instruction in the elementary classroom. In Dreenben R. Thomas JA (eds). **Issues in Microanalysis.** Cambridge: Ballinger, 1980.

CALKINS, D., BORICH, G.D., PASCONE, M., KUGLE, C.L. Generalizability teacher behaviours across classroom observation systems. In Borich, G., et al (eds). **Classroom observation data : Is it valid ? Is it generalizable ?** Austin : University of Texas at Austin, 1977.

ERLICH, O., BORICH, G. Occurence and generalizability of scores on a classroom interaction instrument. **Journal of Educational Measurement** 1979 ; 16(1) : 11-8.

ERLICH, O., SHAVELSON, R.J. The search for correlations between measures of teacher behavior and student achievement : measurement problem, conceptualization problem or both ? **Journal of Educational Measurement** 1978 ; 15(2) : 77-89.

McGAW, B., WARDROP, J.L., BUNDA, M.A. Classroom observation schemes : where are the errors. **American Educational Research Journal** 1972 ; 9(1) : 13-27.

SHAVELSON, R., ATWOOD-RUSSO, N. Generalizability of measures of teacher effectiveness. Educational Research 1977 ; 19(3) : 171-83.

SMITH, P.L. Sampling errors of variance components in small sample multifacet generalizability studies. Journal of Educational Statistics 1978 ; 3(4) : 319-46.

VELLEMAN, P.F., HOAGLIN, D.C. Applications, basics and computing of exploratory data analysis. Boston : Duxbury Press, 1981.